Confidence Interval for a Binomial Proportion Based on Skew-Normal Distribution

Jose A.T. Sanqui[1] and Amanda McGough[2]
[1]Department of Mathematical Sciences, Appalachian State University, 121 Bodenheimer Drive, Boone, NC 28608
[2]Department of Statistics, Virginia Tech, 250 Drillfield Drive, Blacksburg, VA 24061

Abstract
We investigate coverage probabilities and average lengths of confidence intervals for a Binomial proportion based on the skew-normal approximation to the Binomial distribution. We compare the skew-normal based methods with some existing methods such as the Agresti-Coull and the classical normal distribution based methods.

Key Words: Estimation, Confidence Interval, Coverage Probability, Binomial Approximation, Wald's Interval

1. Introduction

1.1 The Skew-Normal Approximation to the Binomial Distribution
The *Skew-Normal* distribution, first investigated by Azzalini (1985) is a fairly recent distribution that includes the standard normal distribution as a special case. In its basic form, its probability density function (p.d.f.) is given by $f(x;\lambda) = 2\varphi(x)\Phi(\lambda x)$ for all real numbers x and λ, with the latter determining the skewness of the distribution and φ and Φ being the p.d.f. and the cumulative distribution function (CDF) of the standard normal distribution, $N(0,1)$. The distributed is denoted as the $SN(\lambda)$ distribution and it's clear that $SN(0) = N(0,1)$. If $X \sim SN(\lambda)$ then $Y = \frac{X-}{\sigma}$ will have a shifted and scaled Skew-Normal distribution with location parameter and scale parameter σ and its distribution is denoted by $SN(\ ,\sigma,\lambda)$. Chang et al (2008) showed that this *Skew-Normal* distribution with the location and scale parameters provides an improved approximation to the $Binomial(n,p)$ distribution compared with the classical normal approximation when the $Binomial(n,p)$ is not symmetric, which is in a way not surprising since the Skew-Normal distribution is also not symmetric when $\lambda \neq 0$.

1.2 Confidence Interval Estimation of a Binomial Proportion p Based on the Skew-Normal Approximation
The improved approximation to the $Binomial(n,p)$ distribution using the $SN(\ ,\sigma,\lambda)$ distribution prompted Chang et al (2008b) to investigate confidence interval methods for estimating the Binomial proportion p based on this approximation similar to the classical confidence interval method based on the normal approximation to the $Binomial(n,p)$ distribution, also known as Wald's interval method. In their paper, Chang et al (2008b) compared the performance of five confidence interval methods: (1) the Wald's method (CI(N)), (2) the Wald's method with continuity correction (CI(N_c)), (3) the Skew-Normal approximation based method (CI(SN)), (4) the Skew-Normal approximation with continuity correction based method (CI(SN_c) and (5) the considered gold standard in estimating the Binomial proportion p, the Clopper-Pearson method also known as the "exact" method (CI(E)) that appeared in Clopper and Pearson (1934). The paper's

general conclusion is that none of the five methods is uniformly superior based on the standard measures of performance, the interval methods' coverage probability (CP) and average length (AL).

1.3 The Agresti-Coull Method

Agresti and Coull (1998) proposed a method for obtaining a confidence interval estimate for the Binomial proportion p and shown that this method has better coverage probability than Wald's method (CI(N)). The Agresti-Coull method (CI(AC)) also known as the adjusted Wald method but better known as the "add two successes and two failures" method is becoming one of the standard recommended methods for estimating p in elementary statistics textbooks. Chang et al (2008b) did not include this method in their study so we want to investigate how does this method compares with the other methods.

1.4 The Agresti-Coull Skew-Normal Approximation Based Method

Similar to the Agresti-Coull method of adjusting the Wald confidence interval by "adding two successes and two failures", a new method for estimating the Binomial proportion p can be obtained by adjusting the Skew-Normal approximation based method using the same adjustment. This new method is what we will be calling the Agresti-Coull Skew Normal approximation based method (CI(ACSN)). In the next section, we will investigate via simulation the performance of this new method compared with the CI(AC) method and two other existing methods investigated in Chang et al (2008b), namely CI(N) and CI(SN).

2. Simulation Design and Results

2.1 Simulation Purpose and Design

Like in Chang et al (2008b), we compared the performance of the four methods, namely, CI(N), CI(SN), CI(AC) and CI(ACSN). We did not include CI(N_c), CI(SN_c) and CI(E) since none of these three methods are shown to be uniformly superior than the CI(N) and CI(SN). We wanted to see primarily how CI(AC) and CI(ACSN) compare with their unadjusted versions CI(N) and CI(SN), respectively.

The factors in our simulation design were the parameters n (50, 100, 150, 200, 500, 1000, 1500 and 2000) and p (.05, .1, .15, …, .95) of the $Binomial(n,p)$ distribution. We randomly generated 10,000 random samples from the $Binomial(n,p)$ distribution for each combination of these two parameters. We then calculated the average length (AL) by averaging the lengths of the 10,000 confidence intervals for each method and the coverage probability by calculating the fraction of the 10,000 confidence intervals that contains p for each of the four methods. We fixed the nominal confidence level to .95. We used R version 3.2.3 software in our simulation.

2.2 Simulation Results

For the simulation described in the previous section, we compared the simulated CP and AL for the four methods: CI(N), CI(SN), CI(AC) and CI(ACSN). Some typical results of the simulation for some combinations of the Binomial parameters n and p are given in Table 1 to Table 4. Table 1 and Table 2 are typical of a large sample result that the performance of the adjusted methods is as expected very similar to the performance of their respective unadjusted versions. The Skew-Normal based methods (adjusted and unadjusted) typically have coverage probabilities that are higher than the nominal level (i.e., CI(ACSN) and CI(SN) are conservative) whereas the Normal based methods have

coverage probabilities that are much closer to the nominal level but can be lower. The average lengths for the normal based methods are typically lower than those for the Skew-Normal based methods. Thus when the sample size is large, the Normal based methods seem to be the preferred methods unless we want to make sure the nominal level is achieved or exceeded.

Table 3 and Table 4 are typical of small to moderate sample size results. From Table 3, we see that CI(ACSN) is the most conservative, closely followed by CI(SN) so again the two Skew-Normal based methods have coverage probabilities higher than the nominal level. The CI(N) method can have coverage probability that is much lower than the nominal level. The CI(AC) typically is the one that has coverage probabilities that are closest to the nominal level although they can be slightly lower. In terms of average length, the two Skew-Normal based methods again are the worst and the CI(AC) is the best. Our new method, the CI(ACSN) has shorter average length than its unadjusted version. Hence when the sample size is small to moderate, it appears the CI(AC) should be the preferred method unless again we require to achieve the nominal level.

Table 1: Coverage Probabilities for *CI(N)*, *CI(SN)*, *CI(AC)* and *CI(ACSN)*

	n=1000, nominal level =.95			
P	*CI(N)*	*CI(SN)*	*CI(AC)*	*CI(ACSN)*
.05	0.9425	0.9759	0.9488	0.9837
.10	0.9517	0.9735	0.9481	0.9760
.15	0.9476	0.9726	0.9562	0.9750
.20	0.9472	0.9651	0.9471	0.9685

Table 2: Average Length for *CI(N)*, *CI(SN)*, *CI(AC)* and *CI(ACSN)*

	n=1000, nominal level =.95			
P	*CI(N)*	*CI(SN)*	*CI(AC)*	*CI(ACSN)*
.05	0.0270	0.0324	0.0274	0.0330
.10	0.0372	0.0430	0.0374	0.0432
.15	0.0442	0.0500	0.0444	0.0500
.20	0.0496	0.0550	0.0496	0.0550

Table 3: Coverage Probabilities for *CI(N)*, *CI(SN)*, *CI(AC)* and *CI(ACSN)*

	n=50, nominal level =.95			
p	*CI(N)*	*CI(SN)*	*CI(AC)*	*CI(ACSN)*
.45	0.9329	0.9765	0.9495	0.9904
.50	0.9449	0.9768	0.9573	0.9850
.55	0.9493	0.9743	0.9504	0.9766
.60	0.9512	0.9745	0.9533	0.9759

Table 4: Average Length for *CI(N)*, *CI(SN)*, *CI(AC)* and *CI(ACSN)*

	n=50, nominal level =.95			
p	*CI(N)*	*CI(SN)*	*CI(AC)*	*CI(ACSN)*
.45	0.2728	0.2976	0.2632	0.2854
.50	0.2744	0.2956	0.2644	0.2834
.55	0.2728	0.2976	0.2632	0.2854
.60	0.2686	0.3012	0.2596	0.2890

3. Conclusions

Our simulation suggests the following conclusions:

- When n is large, the performance of the adjusted methods are very similar to their unadjusted versions. In this scenario, any of the two Normal based methods should be the preferred method for estimating p unless the nominal level is required to be achieved or exceeded.
- When the sample size n is small to moderate, CI(AC) should be the preferred method. CI(N) can have much lower coverage probability than the nominal level. The CI(ACSN) has shorter average length than the CI(SN) but still larger than either of the Normal based methods.
- If we have to choose among the four methods regardless of the sample size, the CI(AC) seem to be the best method unless the nominal level is required to be achieved, in which case, the CI(ACSN) should be used.
- Strictly speaking, none of the four methods we investigated is uniformly superior.
- Since the confidence interval estimation methods we considered can be inverted to obtain hypothesis testing procedures for the Binomial parameter p, the hypothesis testing procedure based on the adjusted Wald method should be the overall best procedure for testing p among the four procedures (i.e., the adjusted Wald hypothesis testing procedure is expected to have generally better power and lower probability of Type I error.

References

Agresti, A. and Coull, B. (1998). “Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions”. *The American Statistician*, Vol. 52, No. 2, pp. 119-126.

Azzalini, A. (1985). "A class of distributions which includes the normal ones". *Scandinavian Journal of Statistics* ***12****: 171–178.*

Chang, C.H.., Lin, J.J., Pal, N. and Chiang, M.C.. (2008). “A Note on the Improved approximation of the Binomial Distribution”.*The American Statistician,* Vol. 62, No. 2, pp. 167-170.

Chang, C.H.., Lin, J.J., Pal, N. and Chiang, M.C. (2008b). “Discussion on Skew-Normal Approximation of a Binomial Distribution”, https://www.researchgate.net/profile/Nabendu_Pal/publications/2.

Clopper, C. J., and Pearson, E. S. (1934), "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," Biometrika, 26, 404-413.